BREAST CANCER STAGE CLASSIFICATION ON DIGITAL MAMMOGRAM IMAGES

Dr. G. Rasitha Banu[1], Fathima N Sakeena [2], Mrs.Mumtaj [3], Mr.Agha Sheraz Hanif [4]

[1]Assistant Professor, Faculty of PHTM,Dept. of HI,Jazan University,KSA
[2] Lecturer,Faculty of CS& IS,Jazan University,KSA
[3] Assistant Professor,Dept.of bioinformatics,Md.Sathak college,India
[4] Lecturer,Faculty of PHTM,Department of HI,Jazan University,KSA

ABSTRACT

Breast cancer is a disease in which the cells of the breast grow out of control, creates an abnormality in the breast tissue. It is the second leading cause of death in women worldwide. In Saudi Arabia, Ministry of health reported that the number of new cases of cancer is 2741 including about 19.9% of breast cancer in women due to unawareness , it usually occurs in women at the age of 52. It accounts for about 22% of all new cancers in women. In developing countries there are still large numbers of breast cancers diagnosed in later stages. So the death rate is also high. To prevent people from this disease, it should be detected at an earlier stage which reduces death rate. Digital mammogram is used for this purpose. The suspected symptoms causing breast cancer are age, post menopause, stress, family history, physical inactivity, obesity, hormonal imbalances and genetically mutated abnormalities. Our work focus on detecting stage of breast cancer using image processing techniques and data mining technique is used to classify the stage of breast cancer and the performance of classifier is evaluated through confusion matrix.

Key Words: Image Processing,Data mining,Weka,Classification,J48.REPTree

1. Introduction

Breast cancer stage is described the condition of cancer, based on its location, its size, where it spreads and the extent of its influence on other organs. In general, the level of breast cancer varies from stage 0 to stage IV. Among various diagnostic techniques, such as X-ray, MRI, breast ultrasound, digital mammograms are the most reliable and inexpensive to detect the symptoms of breast cancer at the early stage, can disclose many information about these abnormalities like masses, micro calcifications, architectural distortion and bilateral asymmetry.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Digital Mammogram is one of the efficient technique to detect the cancer at an earlier stage. There is a special detector which converts a X-ray energy into digital image. It helps the people to reduce the morality rate. It detects abnormalities easily. It is advisable to all women should do regular screening text in the age of 35 to prevent from this disease. There are many advantages of digital mammogram such as: patient spend less time for screening, radiologist quickly transmit the images to another physician and they can be easily manipulated.

Data Mining is a process of discovering hidden patterns in the database. There are many techniques available such as neural networks, association rule mining, classification and clustering and so on.

In our work, we have used data mining tool weka to classify the stage of breast cancer from digital mammographic images.

2. Objective:

1. The main objective of our work is to detect the stage of breast cancer from digital mammographic images based on area of size of the pixel.
2. This computer aided diagnostic system is used is support the radiologist to determine the stage of breast cancer and as an aid in decision making.
3. Classifying the stage of the breast cancer using data mining classification techniques.

3.Proposed Methodology:

Breast masses and micro calcifications are the main indications of abnormalities in digital mammograms. Breast cancer detection can be carried out by using various image processing techniques. The proposed method involves data collection, image preprocessing, segmentation of ROI, feature selection and classification of cancer stages in abnormal mammograms.

1. **Data collection**: Mammography Image Analysis Society (MIAS) database used in this research. Data is in the form of PGM (Portable Gray Map) format. In this research, 50 mammogram images are used for determining the various stages.
2. **Preprocessing**: The noise removal is done by using Gaussian filter. Gaussian smoothing is very effective for removing Gaussian noise, the degree of smoothing is controlled by σ, which is set as 1. The contrast of mammogram image is increased by using Cumulative Histogram Equalization, which has good performance.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

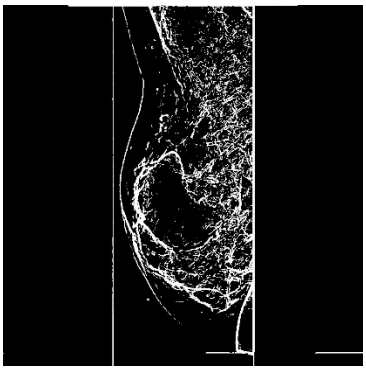


Image1: After preprocessing

3. **Segmentation of ROI**: Segmentation is the process of partitioning a digital image into multiple segments. Segmentation can be carried out by using local thresholding. Edge detection is used to divide into areas corresponding to different objects to enhance the tumor area in mammographic images.
4. **Feature extraction and selection**: Using ROI, the area of size of the pixel can be calculated to identify the various stages of the breast cancer.
5. **Classification:** The process of assigning a label to unknown objects. It is a supervised learning, the image attributes (features) are given as the input to data mining classifiers such as J48 and RepTree to classify the stage of the breast cancer on digital mammograms.

4. Experiments with Weka:

In this research, 50 malignant mammogram images from MIAS database are used, where 16 images are from the group of malignant mammogram dense-glandular, 16 malignant mammogram images derived from fatty group and 18 malignant mammogram images derived from fatty-glandular groups. After the process of preprocessing, segmentation tumor area will be identified. Further using Region of interest the area of pixel can be calculated. Depends on the value of the pixel, the stage of the cancer to be identified. The following table shows Table 1 show the result of determining the stage of cancer from malignant digital mammogram images.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Ref No	Tissue	abnormality	Severity	Radius	area	class
mdb023	G	CIRC	M	29	22268	1
mdb028	F	CIRC	M	56	9385	1
mdb058	D	MISC	M	27	8698	1
mdb072	G	ASYM	M	28	22342	1
mdb075	F	ASYM	M	23	11328	1
mdb090	G	ASYM	M	49	39032	2
mdb092	F	ASYM	M	43	5184	1
mdb095	F	ASYM	M	29	34833	2
mdb102	D	ASYM	M	38	30786	2
mdb105	D	ASYM	M	98	161097	4
mdb110	D	ASYM	M	51	45413	2
mdb111	D	ASYM	M	107	56732	2
mdb115	G	ARCH	M	117	81616	3
mdb117	G	ARCH	M	84	47906	2
mdb120	G	ARCH	M	79	67896	3
mdb124	G	ARCH	M	33	26426	2
mdb125	D	ARCH	M	60	31840	2
mdb130	D	ARCH	M	28	74694	3
mdb134	F	MISC	M	49	6505	1
mdb141	F	CIRC	M	29	63602	3
mdb144	F	MISC	M	27	20944	1
mdb155	F	ARCH	M	95	6957	1
mdb158	F	ARCH	M	88	641	0
mdb170	D	ARCH	M	82	11499	1
mdb171	D	ARCH	M	62	162560	4
mdb178	G	SPIC	M	70	13680	1
mdb179	D	SPIC	M	67	65330	3
mdb181	G	SPIC	M	54	24702	1
mdb184	F	SPIC	M	114	32590	2
mdb186	G	SPIC	M	47	2535	0
mdb202	D	SPIC	M	37	1901	0
mdb206	F	SPIC	M	17	12891	1

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

mdb209	G	CALC	M	87	57756	2
mdb211	G	CALC	M	13	9913	1
mdb213	G	CALC	M	45	5656	1
mdb231	F	CALC	M	44	39429	2
mdb238	F	CALC	M	17	186754	4
mdb239	D	CALC	M	25	156879	4
mdb241	D	CALC	M	38	37691	2
mdb249	D	CALC	M	64	1426	0
mdb253	D	CALC	M	28	58355	2
mdb256	F	CALC	M	37	9141	1
mdb264	G	MISC	M	36	32455	2
mdb265	G	MISC	M	60	66420	3
mdb267	F	MISC	M	56	41947	2
mdb270	G	CIRC	M	72	9738	1
mdb271	F	MISC	M	68	1949	0
mdb274	F	MISC	M	123	11251	1
mdb245	F	CALC	M	38	10734	1
mdb250	D	CALC	M	64	2956	0

Table 1. : The results of determining stage of cancer from Digital Mammogram image.

Out of 50 images, 6 images are belong to stage 0, 19 images are belong to stage I, 15 images come under stage II, 6 images come under stage III and 4 images are belong to stage IV.

The open source software Waikato Environment for knowledge Analysis 3.7(WEKA) is used for our experiment. It is a collection of machine learning algorithms for data mining tasks. Weka can be downloaded from the website [10].

4.1 Performance Measure of Classifiers:

In our experiment, breast cancer data is supplied to classifier of J48, and Random tree algorithms to classify the stages of breast cancer. The classifiers performance are evaluated through Confusion Matrix.

a. Confusion Matrix

It is used for measuring the performance of classifiers. In the confusion matrix, correctly classified instances are calculated by sum of diagonal elements TP (True Positive) and TN (True Negative)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and others as well as FP (false positive) and FN (False Negative) are called incorrectly classified instances.

b. Accuracy

It is defined as the ratio of correctly classified instances to total number of instances in the dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

5. Result Analysis:

There are totally 50 records in the breast cancer dataset. Among these 19 instances belongs to stage 0, 15 instances belongs to stage I, 4 instances belongs to stage II, 6 instances belongs to stage III, 6 instances belongs to stage IV. The following table shows confusion matrix with 12 attributes.

The following Table 2 represents confusion matrix for Random Tree Algorithm

Target class	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	18	0	0	1	0
Stage I	9	3	0	3	0
Stage II	2	2	0	0	0
Stage III	3	2	0	1	0
Stage IV	5	1	0	0	0

Table 2: Confusion matrix for Random Tree Algorithm

In Random tree classifier, the correctly identified instances are 22 and incorrectly identified instances are 28.
The following Table 3 represents confusion matrix for J48Algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Target class	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	18	1	0	0	0
Stage I	0	14	0	1	0
Stage II	0	0	4	0	0
Stage III	0	0	1	5	0
Stage IV	1	0	0	0	5

Table 3: Confusion matrix for J48 Algorithm

In J48 classifier, the correctly identified instances are 46 and incorrectly identified instances are 4.

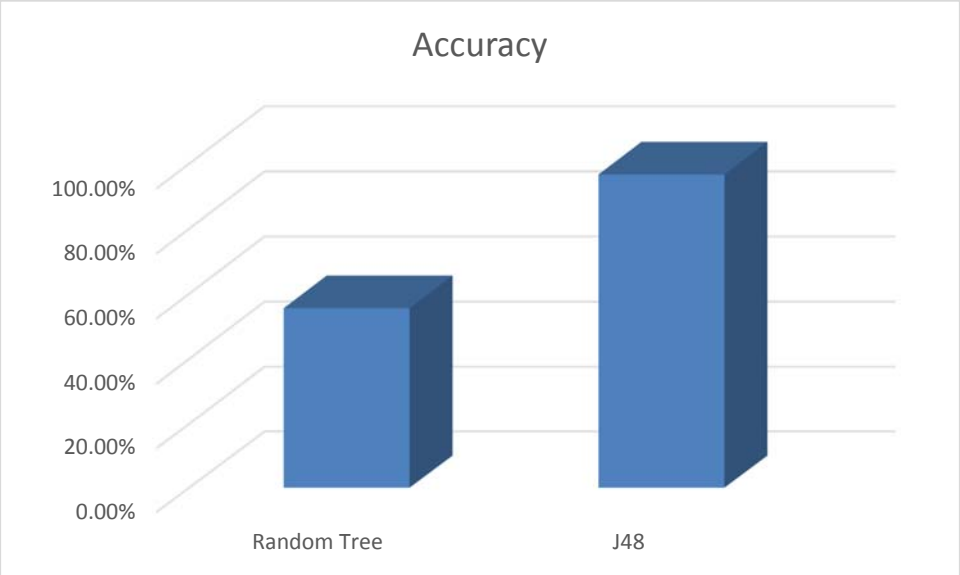
The following Table 4 depicts detailed accuracy of J48, Random Tree algorithm

Classifier	Accuracy
Random Tree	55.55%
J48	96.66%

Table 4: Accuracy of classifiers

Table 4 shows that J48 is giving highest accuracy.

The following chart1 shows the accuracy of classifiers.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Chart 1: Performance Analysis of classifiers

In this chart, X axis represent the algorithm and Y axis represent the accuracy. It shows that the accuracy of J48 is 96.66 % which is best than Random Tree Algorithms.

6. Conclusion

In our research, 50 mammogram images from MIAS database are used. We have used image processing techniques such as Gaussian filtering, histogram equalization, thresholding, edge detection are used to remove the noise, enhance the image, and find the region of interest. The image attributes are extracted from the processed image, according to the area of the size of the pixel, stage of the breast cancer identified. Out of 50 images, 6 images are belong to stage 0, 19 images are belong to stage I, 15 images come under stage II, 6 images come under stage III and 4 images are belong to stage IV. The breast cancer stages are classified using data mining classifier such as J48 and Rep Tree. The performance of the classifiers are evaluated though confusion matrix in terms of accuracy, in which J48 provides good accuracy.

7. References

1. Karmilasari et.al "Sample K-Means Clustering Method for Determining the Stage of Breast Cancer Malignancy Based on Cancer Size on Mammogram Image Basis", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 3, 2014
2. Sezgin, M., and Sankur, B., "Survey over image thresholding techniques and quantitative performance evaluation". Journal of Electronic Imaging, 13 (1): 146–165. 2004.
3. Maitra, I.K., Nag S., Bandyopadhyay S.K., "A Novel Edge Detection Algorithm for Digital Mammogram", International Journal of Information and Communication Technology Research, Vol 2 No.2, February 2012.
4. Kamdi, S.," Image Segmentation and Region Growing Algorithm", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Vol 2 Issue no.1, October 2011.
5. Priya, D.S., and Sarojini, B., "Breast Cancer Detection In Mammogram Images Using Region-Growing And Contour Based Segmentation Techniques", International Journal of Computer & Organization Trends, Vol.3 Issue 8, September 2013.
6. Martins, L.d.O., Junior, G.B., Silva, A.C., Paiva, A.C. and Gattass, M., "Detection of Masses in Digital Mammograms using K-Means and Support Vector Machine", Electronic Letters on Computer Vision and Image Analysis 8(2) : 39-50, 2009.
7. S.P. Meharunnisa et.al " Detection of masses in digital mammograms using SVM", IJCTA,Vol 8 Issue no.3,2015,ppno:899-906.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500